Engineering Applications of Artificial Intelligence 35 (2014) 237–245

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Spectral clustering for sensing urban land use using Twitter activity

Vanessa Frias-Martinez [a], Enrique Frias-Martinez [b,*]

[a] College of Information Studies, University of Maryland, College Park, MD 20742, United States
[b] Telefónica Research, Distrito Telefónica, 28050 Madrid, Spain

ARTICLE INFO

Article history:
Received 6 December 2013
Received in revised form
30 May 2014
Accepted 18 June 2014
Available online 26 July 2014

Keywords:
Urban computing
Crowd behavior
Land use detection
Spectral clustering

ABSTRACT

Individuals generate vast amounts of geolocated content through the use of mobile social media applications. In this context, Twitter has become an important sensor of the interactions between individuals and their environment. Building on this idea, this paper proposes the use of geolocated tweets as a complementary source of information for urban planning applications, focusing on the characterization of land use. The proposed technique uses unsupervised learning and automatically determines land uses in urban areas by clustering geographical regions with similar tweeting activity patterns. Three case studies are presented and validated for Manhattan (NYC), London (UK) and Madrid (Spain) using Twitter activity and land use information provided by the city planning departments. Results indicate that geolocated tweets can be used as a powerful data source for urban planning applications.

1. Introduction

Urban planning is a process that focuses on the control and on the design of urban environments in order to increase the well being of citizens. An important concern in urban planning is the characterization of urban land use, such as residential, industrial or parks. How each part of an urban landscape is used, is mainly determined by zoning regulations. In the context of urban planning, urban zoning is defined as the designation of permitted uses of land based on mapped zones which separate one set of land uses from another (for example residential areas from industrial areas). One of the problems of zoning is to actually evaluate to which extent the areas are being used as required or planned, because the collection of data has to be done on site. Such information is usually gathered through direct observation or using questionnaires that attempt to capture how citizens interact with their urban environment. This traditional approach has some limitations such as the resiliency of citizens to provide answers or the cost of running questionnaires, which highly limits the frequency with which the information is captured. Alternative approaches such as GIS (Geographic Information Systems) (Yin et al., 2011) provide satellite imagery that might reveal some types of land use information through image processing techniques. However, such techniques fail to provide real time information as images are not captured frequently and the land uses that can be identified do not cover the variety of land uses present in a city.

With the increasing capabilities of mobile devices, individuals leave behind footprints of their interaction with urban environments. In this context, cell phones have become one of the main sensors of human behavior, thanks, among others, to their growing penetration and wealth of social applications. As a result, new research areas, such as urban computing and smart cities, focus on improving the quality of life in an urban environment by understanding the city dynamics through the data provided by ubiquitous infrastructures and technologies. New data sources (including GPS, Bluetooth, WiFi hotspots, geo-tagged resources, cell phone traces, etc.) are becoming more relevant for urban planning applications such as transport planning (Frias-Martinez et al., 2012b), traffic estimation (Caceres et al., 2012) or social studies (Oloritun et al., 2013).

In the literature we can find some approaches that use different pervasive infrastructures for the automatic identification of land uses such as GPS (Yuan et al., 2012), cell phone traces (Soto and Frias-Martinez, 2011) or social media applications such as Foursquare (Noulas et al., 2011). In general these approaches tend to focus on specific land uses, on a specific city, and they lack a quantitative validation of the results. Also, GPS and cell phone traces are difficult to obtain due to privacy concerns, and in general, the location information available in social media applications is limited, with the exception of Twitter.

In this paper we propose to use Twitter geolocated data for the automatic identification of land uses. The proposed approach exclusively makes use of spatial (geo-tagged) and temporal

* Corresponding author. Tel.: +34 913379302; fax: +34 913376999.
E-mail address: enrique.friasmartinez@telefonica.com (E. Frias-Martinez).

http://dx.doi.org/10.1016/j.engappai.2014.06.019
0952-1976/

(time-stamped) information of tweets, without accessing personal details or the content of the user-generated information. By doing so, it preserves privacy and can potentially be applied and/or complemented with any other mobile social media dataset with geolocation information. Our novel approach is designed to identify all possible land uses using spectral clustering, it is validated using real land use data provided by city planning departments and is tested in three different urban environments: Manhattan (NYC), London (UK) and Madrid (Spain).

2. Related work

Our work arises as a combination of two research areas mainly crowd modeling and urban computing for urban planning. Different authors have used a variety of user-generated content services for implementing crowd behavior models. Wakamiya et al. (2011) and Fujisaka et al. (2010) studied how to exploit geolocated tweets and the semantics of its content to interpret individual and crowd behavior i.e., how individuals and groups of people move across geographical areas. They propose models of aggregation and dispersion as a proxy to understand the bursty nature of human mobility. Similarly, Kinsella et al. (2011) used geolocated tweets, together with their content, to create geographical models at varying levels of granularity (from zip codes to countries). The authors use these models to predict both the location of the tweet and the user based on location changes.

Recently, in the area of urban computing for urban planning we can find a variety of results using geolocated information to model land use. The approaches can be divided according to its source of information: (1) location-based social networks (LBSN) traces from Foursquare or Twitter; (2) call detail records (CDRs) from cellphones; and (3) GPS traces. The three data sources represent a compromise between granularity and data generality: while GPS data has longitude–latitude information every couple of seconds for usually a very limited number of individuals (usually less than one hundred); CDRs have location information for millions of individuals at a tower level only when an interaction (call or SMS) takes place. LBSN offer an intermediate solution where location information is in the form of longitude–latitude for an intermediate number of individuals (usually hundred of thousands).

Regarding LBSN, Noulas et al. (2011) have used the geolocated information provided by Foursquare to model crowd activity patterns in London and New York City using spectral clustering. For that purpose, the authors characterize the activity patterns identified by the clusters using the predefined Foursquare categories that give an indication of the type of check-in location (restaurants, academic, etc.). As such, this approach gives an approximated understanding of land use. Frias-Martinez et al. (2012a), Neuhaus (2011) presented preliminary results on using Twitter for characterizing urban landscapes. Both studies showed that geolocated tweets potentially contain enough information to identify some land uses, although with some limitations. In the case of Frias-Martinez et al. (2012a) the main limitations arise due to the problems of using *k*-means as clustering technique to classify the terrain. In a related work Cranshaw et al. (2012) presents a new clustering model designed to study land use and social dynamics on a large scale using Foursquare. The results are validated with personal interviews that confirm the clusters identified.

As for CDRs, Calabrese et al. (2011) used cell-phone records from Rome to study the relation between phone activity and commercial land uses using Principal Component Analysis (PCA) to identify the dominant pattern in each area. The results are qualitatively presented and validated and no land use information is actually used. A similar study was done by Ratti et al. (2006) in Milan. Soto and Frias-Martinez (2011) used the tower activity to characterize and cluster similar location patterns and implemented a qualitative evaluation of the results. Toole et al. (2012) also used activity patterns to cluster land uses using a random forest approach in Boston.

GPS data sources are not as commonly used for land use purposes due to the privacy concerns. In general the traces available come from public transport infrastructure such as buses and taxis, which limits its generalization. Yuan et al. (2012) present an initial study using GPS traces from taxis to derive individual mobility and identify regions of different functions using a predefined map of the points of interest of Beijing, i.e., information of the city was used to derive land uses.

In general, the main limitations of the previous approaches are: (1) lack of a formal validation of the results using independent land use data; (2) the studies are presented just for one city, somehow limiting the potential generality of the proposed approach; (3) some data sources (mainly cell phone traces) have strong privacy limitations and (3) in some cases supervised approaches are used, which implies the need of having initial knowledge of the city to derive land uses. Our approach overcomes such limitations by using an unsupervised technique for land use classification that is based on an intermediate data source (both in the sense of number of individuals and privacy) such as Twitter. The approach is validated with external sources and its generality is tested using three cities as examples.

3. Sensing urban land uses using Twitter

The technique we propose for the automatic identification of urban land uses from geolocated tweets has two steps: (1) land segmentation and (2) land use detection. Note that the land uses (such as residential, industrial, parks, etc.) are not defined before hand but identified by the unsupervised learning technique.

3.1. Land segmentation with geolocated data

Given that we want to sense land uses in different urban regions, the first step consists on partitioning the land into different segments, which can then be characterized by its usage pattern. The partitioning of the area considered has to preserve the topological properties of the geolocated tweets, while respecting the actual shape of the geographical area under study. We approached this problem using Self-Organizing Maps (SOM) (Kohonen, 1990).

SOM is a type of neural network trained using unsupervised learning that produces a two-dimensional representation of the training samples. It consists of nodes each one having a weight vector of the same dimension as the input data and a position in the two-dimensional space. Usually the initial weight of the vectors are random values and the initial arrangement of the nodes is a rectangular grid. The procedure for placing a vector from the input data onto the map is to find the node with the smallest distance metric to the data space vector, which in turn updates the weight and the position of the neuron. In our case, the input data are the latitude & longitude pairs that represent the geolocated tweets over a period of time for a specific urban area. As a result, we use a SOM to build a map that segments the urban land into geographical areas with different concentrations of tweets.

Our SOM consists of a collection of N neurons organized in a rectangular grid $[p, q]$, with $N = p \times q$. Since we can choose any initial size $[p, q]$ for the map, our method explores different map sizes and selects as the best land segmentation map the topology